基于 python-casacore 的射电测量集文件生成方法

孙浩民[1] 邓辉*[12] 梅盈[1] 王锋[12]
广州大学物理与电子工程学院/天体物理中心，广东， 广州， 510006
昆明理工大学，云南，昆明，650051
（* 通信作者 广州大学 denghui@gzhu.edu.cn)

摘要

测量集（MeasurementSet，MS）文件是成为射电天文领域的重要存储文件格式，并逐渐成为射电天文数据存储、分析与共享的标准格式，得到越来越多的天文数据处理软件如 CASA、WSCLEAN 等的支持，在 ALMA、LOFAR 等诸多射电望远镜系统中应用。但长期以来，MS 格式在国内应用较少，技术规范文档非常缺乏。本文结合 SKA 工程桥接阶段相关工作需要，对 MS 格式的基本概念、目录结构和字段设计进行了介绍，在此基础上讨论了利用 python-casacore 调用底层 casacore 生成 MS 文件的方法,并将此功能封装到当前 SKA 的算法参考库（Algorithm Reference Library, ARL）。文中给出了利用 ARL 仿真观测生成 MS 文件的实例，并通过 CASA 软件对生成的 MS 文件成像，经过结果比对，验证了 MS 文件的正确性。本文的工作为 SKA 后续的成像实验、观测模拟和文件存储都提供了关键的保障，在满足 SKA 工程桥接阶段工作需要的同时，也为国内外射电天文数据处理工作提供了参考。

关键词：测量集文件；Python-casacore；CASA
中图分类号： TP274

前言

我国的射电天文在近十年来有了飞跃的发展， 21CMA 望远镜[1]阵列带动了国内低频射电天文研究，国内的 FAST 望远镜[2]曾作为世界上最大的射电望远镜平方公里阵（Square Kilometre Array, SKA）的候选方案之一，后来独立发展成世界上最大的单口径望远镜。国内其它射电望远镜阵列包括建于新疆的天籁阵列[3]、建于内蒙古的明安图太阳频谱日像仪(MUSER)[4]以及中国 VLBI 网，这些射电干涉阵列在天文研究、数据处理、经验积累、人才培养等方面都做出了重要贡献，也为中国参加 SKA 奠定了实质性基础。

天文射电数据的存储是射电天文观测的基础要求。长期以来，FITS 文件一直是天文数据存储的标准格式，针对射电数据存储，在 FITS 基础上发展出了 UVFITS 格式与 FITSIDI 格式等。近年来，测量集（MeasurementSet， MS）文件应用越来越广，已经成为射电天文领域的重要存储文件格式。并逐渐成为射电天文数据分析的标准格式，被 CASA（the Common Astronomy Software Applications package, https://casa.nrao.edu/）、WSCLEAN[5]等天文数据处理软件广泛支持。国内测量集格式文件应用相对较少，中、英文资料较少。针对射电数据的需求，国内射电望远镜往往都是根据各自接收机的特点，自行定

*基金项目：国家重点研发计划（2018YFA0404603），国家自然科学基金委员会-中国科学院天文联合基金资助项目(U1831204，U1631129，U1931141)，云南省重点研发计划(2018IA054)，云南省应用基础研究项目(2017FB001, 2018FB103)，国家自然科学基金青年科学基金资助项目（11903009）资助。

[1] 孙浩民，男，硕士研究生，研究方向：天文技术方法，sunhaomin@cnlab.net；邓辉，教授，通信作者，研究方向：天文技术方法，denghui@gzhu.edu.cn。

义相应的原始数据存储格式。如 MUSER 采用裸二进制的方式保存观测文件以大幅度降低存储空间。在需要进行数据交换时，通过格式转换软件改存为 UVFITS 或 FITSIDI 格式。

随着 SKA 桥接阶段的开展，本文作者在参与桥接阶段的工作中负责进行测试与仿真，为了与其它主流天文数据处理工具相对接，不可避免地遇到了如何生成 MS 文件的问题。因此，本文结合实际开发需求，系统讨论了利用 Python 语言和 python-casacore 实现 MS 文件的生成，对后续工作和其它射电望远镜数据存储与共享有一定的参考作用。

一、测量集文件的概念与定义

1.1 MeasurementSet 文件的基本概念

测量集文件是一个遵从射电干涉测量方程（Radio Interferometer Measurement Equation,RIME[6]）的文件格式，能够保存校准前的射电天文数据。测量集文件的设计标准发布后，CASA 团队和欧洲 VLBI 网等团队的多个天文软件开发小组进行了代码实现。由于 CASA 采用测量方程作为其基本校准方案，MS 文件很自然地成为 CASA 软件射电观测数据的存储标准。随着 CASA 成为 ALMA 和 VLA 的指定数据处理分析软件包[7]，MS 也自然地成为 ALMA 和 VLA 数据分析中的缺省数据格式。ALMA 和 VLA 原始数据存储格式为 ASDM 和 SDM，因此也都开发了相关软件，实现从 ASDM/SDM 数据格式转换为 MS 格式。

MS 文件实际上是一个关系数据库，其格式涵盖了射电天文学中所有可以想到的用例，无论是单碟、还是简单的几个天线组成的射电干涉仪，乃至数百上千天线的大型射电干涉仪。

MS 借鉴了关系型数据库的建模方法来降低数据的冗余，即构建主键与外键的方式，把多次出现的数据（如天线数据等）放入单独的数据库表（即子表）中，然后在数据库的主体部分（即主表）中进行建立相应的索引（主键），实现对子表数据的引用（外键）。在有两层子表的情况下，第一层由主表引用，第二层由第一层引用，即可以通过子表引用其他子表。

MS 文件生成时，大部分数据，包括干涉得到的可见度函数和/或单天线总功率测量值及其时间戳会保存在主表，而大部分元数据(Meta Data)存放在二级子表中。主表中一般需要包括的列是 DATA（干涉数据）或 FLOAT_DATA（纯单碟数据）列，根据不同的射电望远镜，这两列中的一列必须存在。

1.2 MS v2.0 文件结构

CASA 软件中采用 MeasurementSet 第二版本[8]作为数据格式标准。实际， MeasurementSet 集在 AIPS ++ Note 191[2]中就被正式定义。在本文中，为了确保与 CASA 软件的数据兼容性，也完成采用了 MeasurementSet 第二版本为基础标准。以下分析与讨论均按该版本规范开展。

1.2.1 子表结构

CASA 使用的 MeasurementSet 的表格结构见表 1。每个 MS 文件，一定要有一个主表（MAIN）表，表中包含许多数据列和各种子表中的键。 每个子表中至多有一个。 子表存储为 MS 的关键字，所有定义的子表在下表中列出。 可选子表以斜体和括号显示。在实际应用中，非可选子表一定要在数据生成时生成，但部分子表可以没有内容。

表 1. MS V2 版本的表格结构

[2] https://casacore.github.io/casacore-notes/191.pdf

Table.1 Table structure for MS V2 version

子表名称		
子表名	内容	键值(key)
ANTENNA	天线信息	ANTENNA_ID
DATA_DESCRIPTION	数据描述	DATA_DESC_ID
(*DOPPLER*)	多普勒跟踪	DOPPLER_ID, SOURCE_ID
FEED	馈源	FEED_ID, ANTENNA_ID, TIME, SPECTRAL_WINDOW_ID
FIELD	场位置	FIELD_ID
FLAG_CMD	标注命令	TIME
(*FREQ_OFFSET*)	频率偏移	FEED_ID, ANTENNA*n*, FEED_ID, TIME, SPECTRAL_WINDOW_ID
HISTORY	历史信息	OBSERVATION_ID, TIME
OBSERVATION	观测地点与计划	OBSERVATION_ID
POINTING	指向信息	ANTENNA_ID, TIME
POLARIZATION	极化设置	POLARIZATION_ID
PROCESSOR	处理器信息	PROCESSOR_ID
(*SOURCE*)	源信息	SOURCE_ID, SPECTRAL_WINDOW_ID, TIME
SPECTRAL_WINDOW	谱窗	SPECTRAL_WINDOW_ID
STATE	状态信息	STATE_ID
(*SYSCAL*)	系统校准	FEED_ID, ANTENNA_ID, TIME, SPECTRAL_WINDOW_ID
(*WEATHER*)	各天线气象信息	ANTENNA_ID, TIME

显然，与 FITS 文件相比，MS 的格式要复杂很多。在实际应用中，根据射电望远镜的不同观测数据需求，生成相应的子表并存入相关的数据，最终构成一个完整的 MS 文件目录树结构，原则上只要是非可选表都需要生成。限于篇幅的限制，我们在下节中以主表的结构来分析 MS 格式的相关字段要求与数据类型。其它表的结构直接查看 MS 技术规范。

1.2.2 MAIN table: Data, Coordinates and Flags

主表(MAIN TABLE)是 MS 文件中必须存在的一个表。在数据存储上，MS 格式与 FITS 格式基本类似，需要保存的数据类型也包括整型（int)，浮点(float)，双精度(double)、字符串(string)等。

与 FITS 文件的头定义相类似，MS 文件中，每个表有相应的字段设计的概念，但更为复杂， MS 的字段设计分为三种情况：

1、 关键字：MS_VERSION，用来标识所保存的 MS 文件遵从哪一个版本的规范。

2、键(key)：相当于关系型数据库当中的主键，用来和子表进行关联。如 TIME 给出了观测时刻。这些值的定义与写入应完全按照 MS 格式的要求来进行。

3、非键属性：根据实际需要定义的某些重要参数或属性。

其它表的存储与主表的存储完全类似，也有各自规定的相应保留字，键值和参数。要成功生成 MS 格式，本质就是要明确各个字段的属性与格式要求，写入正确的数值并给出正确的单位。

1.3 MS 文件存放结构

一般关系型数据库的存储都只有库文件与索引文件两大关键部分。但 MS 格式有显著的不同，它采用多级目录多文件保存的方法，各个表都以 CASA 表格式存储。这意味着一个表格包含了多个文件，且整个 MS 文件也不是单个文件,而是一个目录树。

MS 文件的目录结构可以看作由多级目录构成。一般来说，主表位于第一级目录下，各个子表位于第二级目录下。每一级目录下，均包含实际数据存放位置信息的 table.info, table.f0, table.f1 等文件。

二、基于 Python-Casacore 的 MS 生成方法

众所周知，从 AIPS++发展到 CASA 后，CASA 软件的开发采用了多计算机语言混合编程的方法，其主要代码来自于 C/C++，然后主要的用户接口与调用部分基本上全部采用 Python 来实现，C/C++部分合成了 casacore 软件，也就是所谓的 CASA 核心库。可以这么说，casacore 是目前最为完整的射电天文数据处理软件包，也是唯一实现了 MS 文件的读写操作的软件包。

为了能够实现在 Python 语言中对 casacore 功能的调用，python-casacore[3]被开发出来，用以实现对 casacore 核心函数库的调用，虽然也有相应的文档介绍，但如何实现 MS 格式的写入并没有详细的文档。为此，在大量实验与验证的基础上，我们分析了 Python-casacore 的基本函数调用，并对这些函数中操作数据表的函数以及参数类型进行了分析。在此基础上，我们进一步封装完成了一个完整的 MS 格式输出对象，并将代码集成到了 ARL 中，通过简单地调用这一对象实例就可以完整地实现 MS 文件生成。

2.1 Python-casacore 的表操作

python-casacore 是 casacore 的 Python 封装接口，针对 MS 文件的数据操作，主要提供了如下函数：

表 2. python-casacore 中的相关函数
Table 2. The MS related functions of python-casacore

序号	对象/函数名	基本功能
1	Table	打开 MS 数据表文件
2	putinfo	存放信息
3	putkeyword	存放数据字段
4	putcol	存放一列数据
5	putcell	存放一个数据
6	tableutil.makearrcoldesc	定义列的描述，数据为数组形式
7	tableutil.makescacoldesc	定义列字段，数据为单一数据
8	tableutil.maketabdesc	创建表结构
9	close	关闭文件

2.2 数据子表生成

[3] https://github.com/casacore/python-casacore.

利用表 2 中的方法，可以生成一个完整的子表，在表 3 中给出了一个生成 MS 文件实例的程序流程图，用以说明如何应用 python-casacore 来生成相应的表和填入数据。

表 3 代码段示意与说明

Table 3. The demonstration codes f

编号	程序代码	说明
1	col1 = tableutil.makearrcoldesc('UVW', 0.0, 1, comment='Vector with uvw coordinates (in meters)', keywords={'QuantumUnits': ['m', 'm', 'm'], 'MEASINFO': {'type': 'uvw', 'Ref': 'ITRF'}})	创建 col1 列，定义字段为 uvw，存值为浮点类型，本列一行需要存放数组
2	col2 = tableutil.makearrcoldesc('FLAG', False, 2, comment='The data flags, array of bools with same shape as data')	创建 col2 列，字段为 FLAG，类型为布尔
3	desc = tableutil.maketabdesc([col1, col2, col6, col7])	创建表字段定义 desc
4	tb = table("%s" % self.basename, desc, nrow=0, ack=False)	创建表，按 desc 生成字段
5	for j in range(nBand): fg = numpy.zeros((nBL, self.nStokes, self.nchan), dtype=numpy.bool) tb.putcol('UVW', uvwList, i, nBL) tb.putcol('FLAG', fg.transpose(0, 2, 1), i, nBL)	写入数据，重复次数为波段个数

三、MS 文件生成

算法参考库(ARL)[4]是由射电天文科学家 Tim Cornwell 领头开发的射电干涉阵数据处理算法验证库，用以为后续 SKA 的数据处理提供算法验证。

在参与桥接工作期间，我们根据 ARL 开发的需要，以 ARL 为基础进发了 MS 文件输出模块，用以保存 ARL 数据仿真后的结果，并实现与其它常用天文数据处理软件的数据共享，验证 ARL 生成结果。最终封装后的 MSL 文件类为 WriteMS，类图如图 1 所示。

[4] https://github.com/SKA-ScienceDataProcessor/algorithm-reference-library

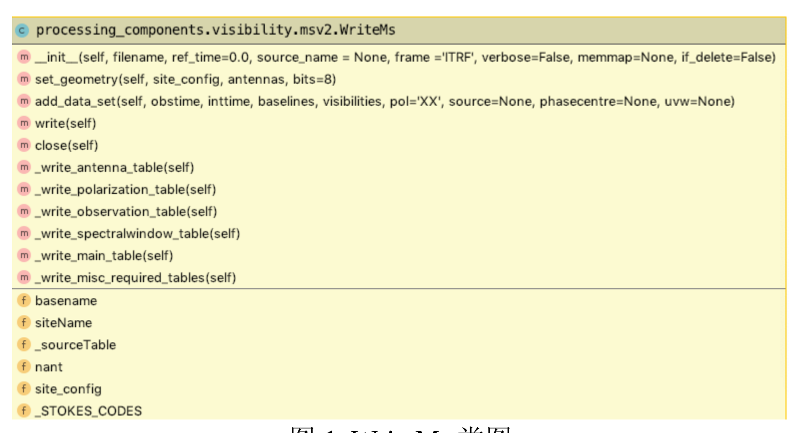


图 1. WriteMs 类图

Figure 1. The class diagram of WriteMs

在实际应用中，用户只需要简单地 import 软件包，然后传入相应的可见度函数数据，最终就可以生成完整的 MS 文件，在 ARL 的代码中我们给出了一个完整的实现代码（test_export_ms_arl.py），下面图 2 中用一个模拟生成 MS 文件实例的程序框图来对 MS 文件的写入和生成进行说明。

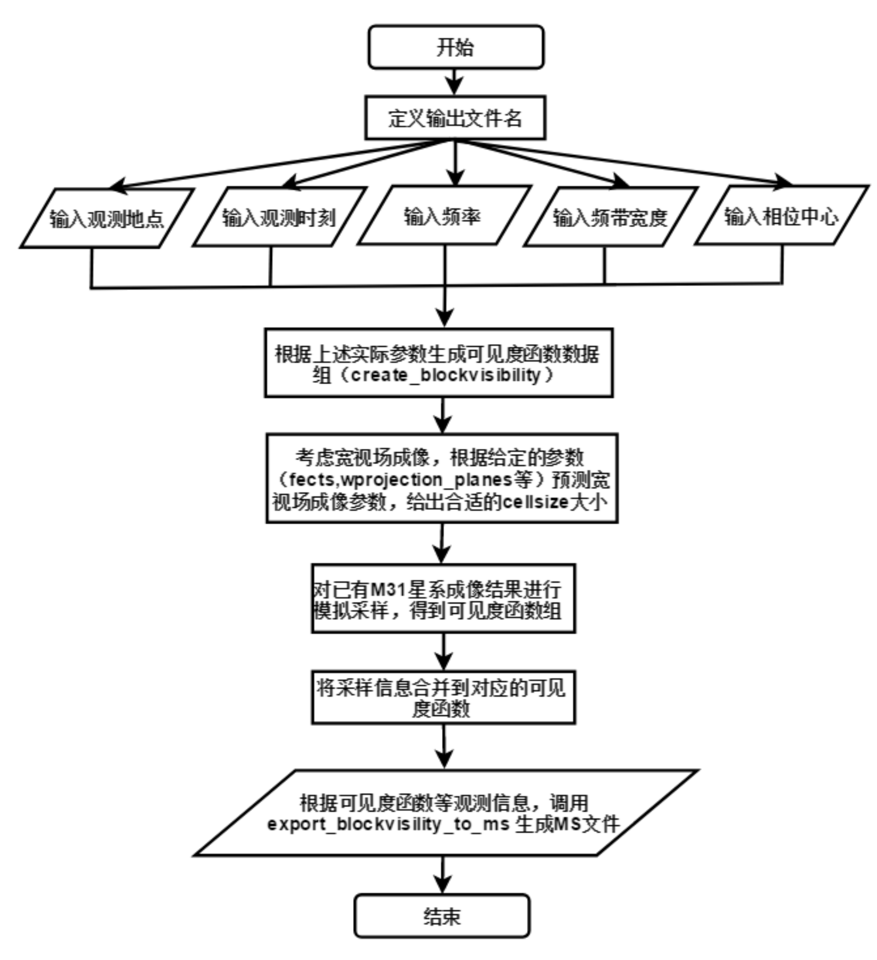


图 2. MS 格式文件生成实例流程图

Figure2. MS format file generation example flow chart

四、结果测试与验证

为了验证 MS 文件是否写入正确，最简单的方法就是先对原本的 MS 格式观测数据进行读取，再用本文中的方法重新将数据写入生成一个新的 MS 格式文件，利用 CASA 软件直接对新生成的 MS 数据做成像处理，并将成像结果与原本的观测结果作对比。因为 CASA 软件对 MS 数据在读取中有较多的校验操作，因此通过了 CASA 的处理就意味着 MS 的各个子表和字段是合理的。虽然在部分字段上的取值可能与真实情况有区别，但这不会影响成像的处理。具体操作如下表所示：

表 4. CASA 操作命令

Table 4. The commands for imaging in CASA

```
CASA <3>: default tclean
CASA <4>: vis='/Users/f.wang/dev/algorithm-reference-library/data/vis/Test_outpu: t.ms'
CASA <5>: cell='74arcsec'
CASA <6>: imsize=512
CASA <7>: imagename='Test_output'
CASA <8>: niter=0
CASA <9>: go
--------> go()
Executing:  tclean()
```

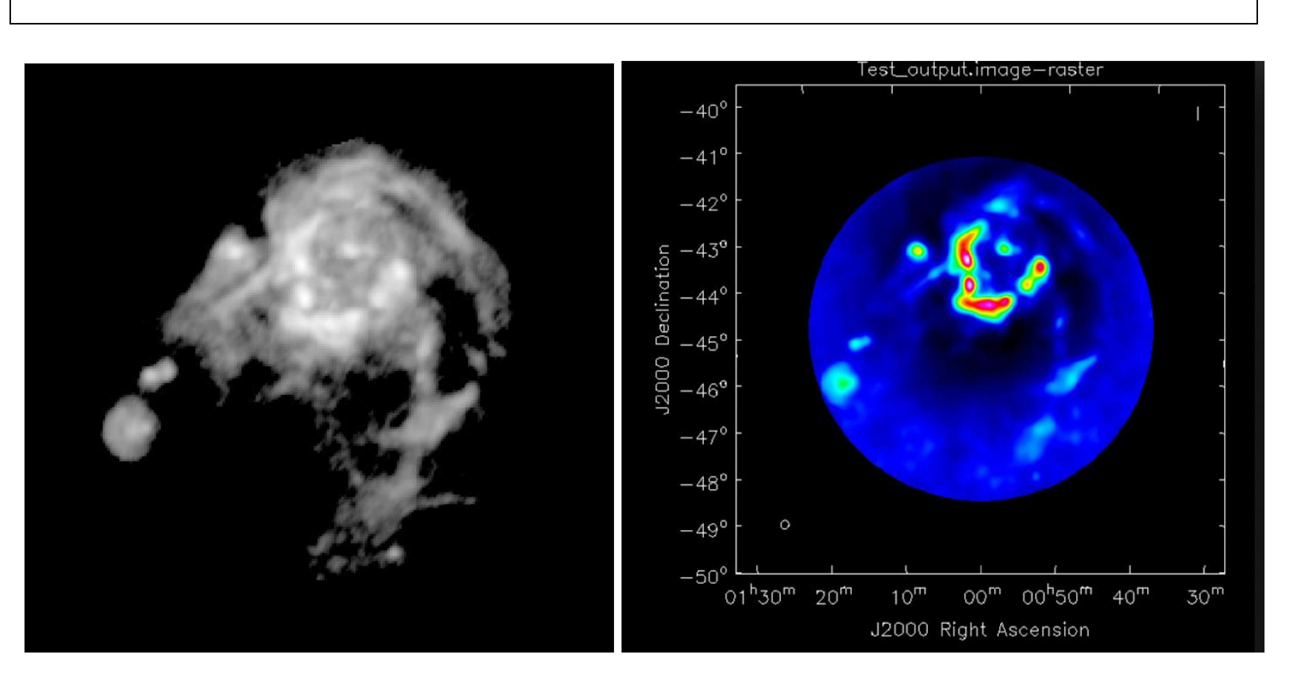


图 3. 左图为 M31 的观测图像，右边为利用 CASA 处理对应的 MS 文件所获得的结果

Figure 3. shows the observation image of M31 on the left and the result of processing the corresponding MS file with CASA on the right.

图 3 给出了 CASA 读取 MS 文件并进行成像得到的脏图（右图）与模拟观测所用的图像（左图），对比发现，基于本文得到的 MS 文件成像结果与实际图像相一致，验证了 MS 文件的正确性。

五、结论

虽然测量集文件规范制定较早，但在我国的射电领域应用较少。一方面是因为 MS 文件会占用较多的空间，另一方面是生成 MS 文件一直依赖于 casacore 这一底层软件包，开发比较困难。因此，本文结合 SKA 工程建设的需要，系统地研究与讨论了 MS 文件的定义、结构与字段设计以及利用 Python-casacore 对数据的写入。同时，最后的实验表明了本文所生成的 MS 文件内容的正确性。目前所有相关代码已经集成到 ARL 软件中，所有的源代码均可以在 https://github.com/SKA-ScienceDataProcessor/algorithm-reference-library 中下载。整体来看，本文的工作在 ARL 的开发过程中起到了关键的作用，为后续的 SKA 数据模拟与文件存储提供了保障，也对其他射电天文数据的 MS 文件生成有较好的参考作用。

六、参考文献

[1] Wu X. Probing the epoch of reionization with 21CMA: status and prospects[C]. Bulletin of the American Astronomical Society, 2009: 474.
[2] 南仁东. 500m 球反射面射电望远镜 FAST[J]. 中国科学 G 辑:物理学、力学、天文学, 2005, (05): 3-20.
[3] 陈学雷. 暗能量的射电探测——天籁计划简介[J]. 中国科学:物理学 力学 天文学, 2011, 41(12): 1358-1366.
[4] Yan Y, Zhang J, Huang G. On the Chinese spectral radioheliograph (CSRH) project in cm-and dm-wave range[C]. 2004 Asia-Pacific Radio Science Conference, 2004. Proceedings., 2004: 391-392.
[5] Offringa A, Mckinley B, Hurley-Walker N, et al. WSCLEAN: an implementation of a fast, generic wide-field imager for radio astronomy[J]. Monthly Notices of the Royal Astronomical Society, 2014, 444(1): 606-619.
[6] Hamaker J, Bregman J, Sault R. Understanding radio polarimetry. I. Mathematical foundations[J]. Astronomy and Astrophysics Supplement Series, 1996, 117(1): 137-147.
[7] Petry D. Analysing ALMA data with CASA[J]. arXiv preprint arXiv:1201.3454, 2012.
[8] Kemball A, Wieringa M. MeasurementSet definition version 2.0[J]. URL: http://casa.nrao. edu/Memos/229. html, 2000.

MeasurementSet File Export Method Based on Python-casacore

Haomin Sun[1], Hui Deng[12*], Ying Mei[1], Shoulin Wei[2], Wei Dai[2], Feng Wang[12]
1. Center for astrophysics,Guangzhou university, Guang Dong, Guang Zhou, 510006
2. Kunming University of Science and Technology, Kunming, Yunnan , 650051
(*corresponding author Guangzhou university denghui@gzhu.edu.cn)

Abstract
Measurement Set (MS) file has been an important storage file format and is gradually becoming the standard format for data analysis of radio astronomy. More and more astronomical data processing software such as CASA and WSCLEAN fully support the MS standard, which lead to MS standard had been widely used in data storage for almost all modern radio telescopes. However the MS format was rarely used in China because of the lack of related technical documents. In this paper, we introduce the basic concept of MS format including directory structure and field design. A method of generating a comprehensive MS file by using Python-casacore package is further discussed in detail. Meanwhile, an example of generating MS file with the visibility functions generated by ARL simulation observation is given. The imaging result using CASA shows that the method presented in this study is correct and can meet the practical application requirements. This method not only satisfies the requirements of the SKAO during engineering bridging stage, but also provides a good reference to the community of radio astronomy data processing in China.

Keywords: Measurement Set (MS) files; Python-casacore